Mining Historical Advertisements in Digitised Newspapers

Wevers, M.

DOI
[10.1515/9783110729214-011](10.1515/9783110729214-011)

Publication date
2023

Document Version
Final published version

Published in
Digitised Newspapers – A New Eldorado for Historians?

License
CC BY-NC-ND

Link to publication

Citation for published version (APA):
Wevers, M. (2023). Mining Historical Advertisements in Digitised Newspapers. In E. Bunout, M. Ehrmann, & F. Clavert (Eds.), *Digitised Newspapers – A New Eldorado for Historians?: Reflections on tools, methods and epistemology* (pp. 227-252). (Studies in Digital History and Hermeneutics; Vol. 3). De Gruyter Oldenbourg. Advance online publication. https://doi.org/10.1515/9783110729214-011

Melvin Wevers

Mining Historical Advertisements in Digitised Newspapers

Abstract: Historians have turned their focus to newspaper articles as a proxy of public discourse, while advertisements remain an understudied source of digitized information. This paper shows how historians can use computational methods to work with extensive collections of advertisements. Firstly, this chapter analyzes metadata to better understand the different types of advertisements, which come in a wide range of shapes and sizes. Information on the size and position of advertisements can be used to construct particular subsets of advertisements. Secondly, this chapter describes how textual information can be extracted from historical advertisements, which can subsequently be used for a historical analysis of trends and particularities. For this purpose, we present a case study based on cigarette advertisements.

Keywords: historical advertisements, text mining, digitized newspapers, digital history

1 Introduction

In recent years, we have seen an explosive growth of digitized historical newspapers. Innovations in natural language processing have extended the possibilities for historians to extract information from large corpora of digitized historical texts. National libraries and projects such as *impresso*, *Newseye,* and *Oceanic Exchanges* offer access to digitized archives of historical newspapers.[1]

For historians, newspapers provide a longitudinal understanding of public discourse.[2] Newspapers are not the only gateway to public discourse, since they do not capture public discourse in its entirety. In its function as a proxy, a newspaper operates as a transceiver; it is both the producer and the messenger of

---

---

public discourse.[3] On a surface level, newspapers inform us about the views of journalists and people that were interviewed by these journalists. However, as Marshall claims, scholars can also uncover the "values, assumptions, and concerns, and ways of thinking that were a part of the public discourse of that time" by analyzing "the arguments, language, the discourse practices that inhabit the pages of public magazines, newspapers, and early professional journals."[4] With easier access to (and increased availability of) newspaper repositories, studies of the representation of ideas, values, and practices in public discourse have gained traction.[5]

Even though newspapers contain a considerable amount of advertisements, these remain an understudied source in computational studies of public discourse. This is surprising since advertisements are rich and varied carriers of information of the past. In his seminal work *Advertising the American Dream*, Marchand argues that adverts offer a lens on the past and provide "insight into the ideals and aspirations of past realities [...] they show the state of technology, the social functions of products, and provide information on the society in which a product was sold".[6] Others point out that, even though adverts provide perspectives on the past, this is a distorted one, as the content of ads is driven by commercial interest.[7] Marchand acknowledges this criticism and conceptualizes advertisements as distorted mirrors. He argues that despite the primary function of advertisements to sell products, they still communicate social and cultural values, albeit in a somewhat distorted manner.[8] Moreover, one could argue that, for ads to be successful, they need to resonate with their audience. As such, they have to be reflective of

3 Michael Schudson (1982). *The Power of News*. Cambridge: Harvard University Press, pp. 17–18.
4 Margaret Marshall (1995). *Contesting Cultural Rhetorics: Public Discourse and Education, 1890–1900*. Ann Arbor: University of Michigan Press, p. 8.
5 J. van Eijnatten and Ruben Ros (2019). "The Eurocentric Fallacy: A Digital-Historical Approach to the Concepts of 'Modernity', 'Civilization' and 'Europe' (1840–1990)." In: *International Journal for History, Culture and Modernity* 7. DOI: https://doi.org/10.18352/hcm.580; Joke Daems et al. (June 2019). "'Workers of theWorld'? A Digital Approach to Classify the International Scope of Belgian Socialist Newspapers, 1885–1940." In: *Journal of European Periodical Studies* 4.1, pp. 99–114. DOI: 10.21825/jeps.v4i1.10187.
6 Roland Marchand (1985). *Advertising the American Dream: Making Way for Modernity, 1920–1940*. Berkeley: University of California Press.
7 Stephen R. Fox (1997). *The Mirror Makers: A History of American Advertising and Its Creators*. English. Urbana: University of Illinois Press.
8 Roland Marchand (1985). *Advertising the American Dream: Making Way for Modernity, 1920–1940*. Berkeley: University of California Press; T. Jackson Lears (1994). *Fables of Abundance: A Cultural History of Advertising in America*. New York: Basic Books.

aspects of public discourse. Still, it is crucial to remain critical of the skewed representation of ideas and values in historical advertisements.

Advertisements are a fascinating and complex historical source, partly because of their multi-modal nature; they contain both visual and textual content. Since there often is an interplay between the visual and textual material, an analysis of only the textual content is somewhat limited. In recent years, advances in computer vision research have provided methods to examine digitized visual material at scale. While the use of computer vision is rapidly evolving and already offers promising methods of analysis, discussion of it falls outside of the scope of this chapter.[9] In this chapter, the topic of concern is the analysis of metadata and textual content.

The ability to extract information from large numbers of advertisements, both synchronically and diachronically, allows scholars to study cultural expressions on a macro-scale. This process has also been described as distant reading.[10] In this chapter, I refrain from using this much-debated term often used in a binary opposition to close reading, where quantitative approaches are incorrectly equated with distant reading and qualitative ones with a form of close reading. When switching between perspectives of analysis, there is more than merely the binary of close and distant. In my view, the use of computational methods can also offer fine-grained contextualized interpretations of particular expressions. More often than not, historical interpretations drawn from advertisements are based on a small selection of ads, which opens the door to cherry-picking. The ability to chart trends over time makes it possible to model whether findings in smaller subsets can be generalized, or whether specific expressions deviate from general patterns. On an intermediate, meso-scale, one can also more easily find variations of a single cultural expression.

While computational methods open up new modes of analysis of historical sources, we have to take into account that the quality of these digitized materials is often sub-optimal, despite continuous advances in Optical Character Recognition (OCR) software and natural language processing. It is especially challenging

9 Melvin Wevers and Thomas Smits (Apr. 2020). "The Visual Digital Turn: Using Neural Networks to Study Historical Images." In: *Digital Scholarship in the Humanities* 35.1, pp. 194–207. DOI: 10.1093/llc/fqy085; Taylor Arnold and Lauren Tilton (Mar. 2019). "Distant viewing: analyzing large visual corpora." In: *Digital Scholarship in the Humanities*. doi: 10.1093/digitalsh/fqz013. eprint: https://academic. oup.com/dsh/advance-article-pdf/doi/10.1093/digitalsh/fqz013/28082598/fqz013.pdf. URL: https://doi.org/10.1093/digitalsh/fqz013; Joon Son Chung et al. (2015). "Re-Presentationsof Art Collections." In: *Computer Vision – ECCV 2014 Workshops*. Ed. by Lourdes Agapito, Michael M. Bronstein, and Carsten Rother. Cham: Springer International Publishing, pp. 85–100.
10 Franco Moretti (2013). *Distant Reading*. London: Verso.

for OCR software to recognize text in advertisements correctly. In addition to more generic factors such as the quality of paper or the printing technique, text in advertisements is often of varying sizes and fonts, or part of a heavily-stylized logo. OCR software regularly turns these forms of textual content into gibberish. The sub-optimal text quality of original documents makes it more challenging to study adverts than articles.

Despite these shortcomings, we can still extract meaningful information and patterns from the OCR-ed text. In this process, we have to be selective of the methods that we use, since their performance can be impacted in different ways by imperfect OCR.[11] In addition to the OCR-ed text, we can also learn about trends in advertising from metadata on the position and size of advertisements in newspapers. We can, thus, study multiple aspects of advertisements using computational means.

This chapter showcases how we can use computational techniques to study advertisements at scale in digitized newspapers. The first section shows how we can use metadata to examine trends in advertising. The second section gives examples of how text mining can be used to extract information from advertisements. This step is explained by demonstrating a case on product nationalities associated with cigarettes. In this case study, we rely on text mining to better understand changes and continuities in the associations with nationalities in advertisements for cigarettes.

2 Metadata Analysis: Advertisements in Almost all Shapes and Sizes

The digitization of newspapers has made it possible to study advertisements at scale using keyword searches. However, not all digitized newspaper collections have segmented the articles and advertisements, presenting users with full-page scans that contain multiple document types, for example, advertisements and articles. In cases where text blocks were segmented but not classified, text

---

11 Daniel van Strien et al. (2020). "Assessing the Impact of OCR Quality on Downstream NLP Tasks:" In: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*. Malta: SCITEPRESS, pp. 484–496; Mark J. Hill and Simon Hengchen (Dec. 2019). "Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study." In: *DigitalScholarship in the Humanities* 34.4, pp. 825–843.

classifiers can be used to identify document types.[12] After training a text classifier with annotated data, the classifier can then predict metadata for unseen new data. Recently, there have been efforts to rely on computer vision to segment and classify the elements on a newspaper page (e.g. articles, tables, ads).[13]

Fortunately, *Delpher*, the digitized newspaper portal hosted by the National Library of the Netherlands (KB), includes segmented documents and metadata on the document type.[14] Access to this type of metadata allows researchers to filter for advertisements. For the twentieth century alone, we can quickly assess that *Delpher* holds over thirty million advertisements in national and regional newspapers. Browsing through the results, it quickly becomes clear that advertisements come in a wide range of shapes and sizes. We can find tiny, square-shaped advertisements; column-shaped classifieds advertisements; and full-page spreads (see Fig. 1 for an example of a column of classified ads). The content and target audiences of these ads vary considerably. How can we deal with this variation, without more specific metadata information?

This section explains how we can use metadata on the size and position of advertisements to cluster types of advertisements and subsequently filter out subsets. The analysis in this section is based on the metadata of the national newspaper *Trouw* (1946–1995).[15] Initial exploratory data analysis shows that this title contains instances of bad segmentation, in which small parts of advertisements appear as separate advertisements. To exclude these segmentation errors, we filtered out ads with a width or height smaller than 100 pixels. After removing

---

12 Aysenur Bilgin et al. (Oct. 2018). "Utilizing a transparency-driven environment toward trusted automatic genre classification: A case study in journalism history." In: *IEEE 14th International Conference on eScience, e-Science 2018*, pp. 486–496. DOI: 10.1109/eScience. 2018.00137.
13 Raphaël Barman et al. (Jan. 2021). "Combining Visual and Textual Features forSemantic Segmentation of Historical Newspapers." In: *Journal of Data Mining & DigitalHumanities* HistoInformatics. URL: https://jdmdh.episciences.org/7097; Benjamin Charles Germain Lee et al. (2020). "The Newspaper Navigator Dataset: Extracting Headlinesand Visual Content from 16 Million Historic Newspaper Pages in Chronicling America." In: *Proceedings of the 29th ACM International Conference on Information Knowledge Management*. CIKM '20. Virtual Event, Ireland: Association for Computing Machinery, pp. 3055–3062. DOI: 10.1145/3340531.3412767. URL: https://doi.org/10.1145/3340531.3412767; Bernhard Liebl and Manuel Burghardt (2020). "From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline." In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, The Netherlands, November 18–20, 2020*. Ed. by Folgert Karsdorp et al. Vol. 2723. CEURWorkshop Proceedings. CEUR-WS.org, pp. 351–373. URL: http://ceur-ws.org/Vol-2723/long20.pdf.
14 http://www.delpher.nl. The metadata on document type was added manually.
15 For most of the examples in this chapter, direct access to the metadata and newspaper data is required. Researchers can contact KB for API access to the newspapers. Code is available on: https://github.com/melvinwevers/eldorado.

Fig. 1: Classified adverts from *De Tijd* (19 January, 1957)
Copyright: De Tijd.

these advertisements, *Trouw* contains about 1.31 million advertisements for the period 1945–1995. From Fig. 2, we can gauge that the total number of advertisements increased between 1946 and 1995, with a sudden increase in the early 1980s. Alongside this increase in the number of ads, we also see a sudden, albeit slight decrease in ads' size (Fig. 3).[16]
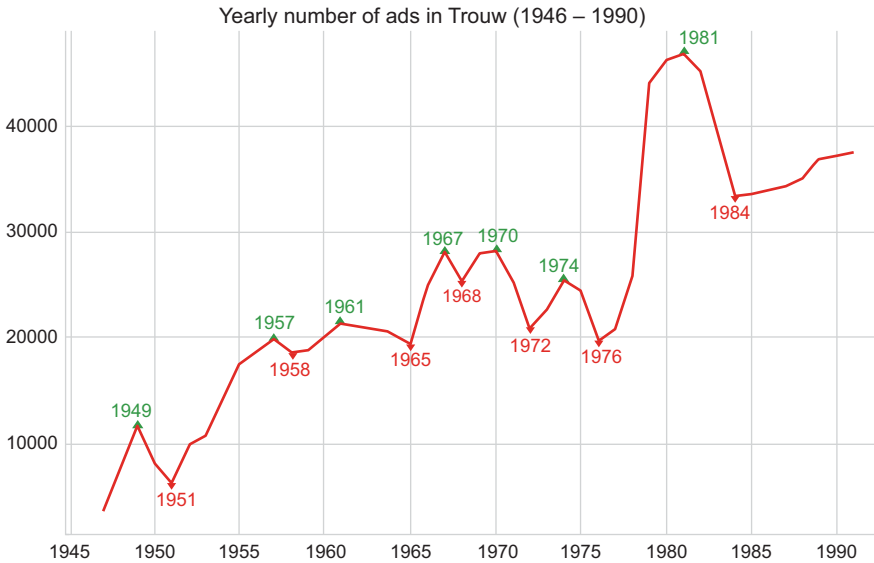

Fig 2: Yearly number of advertisements in *Trouw*. Green triangles indicate peaks and red triangle troughs.

In addition to their size, we can also examine where ads are positioned in the newspaper issue and where they can be found on the individual pages. Using the width and height information of advertisements, we created a heat map that shows which parts of the page are taken up by advertisements. There is a clear difference in pixel density between odd and even pages in *Trouw* (Fig. 4a and 4b). This can be explained by the fact that advertising on odd pages is more expensive than on even pages, as well as the common practice that new sections generally start on the odd pages. Moreover, we see that on the even pages, the upper-left corner, which is closer to the margin, is less populated than the lower-right side.

---

16 We log-transformed the mean size to visualize the rate of change better.
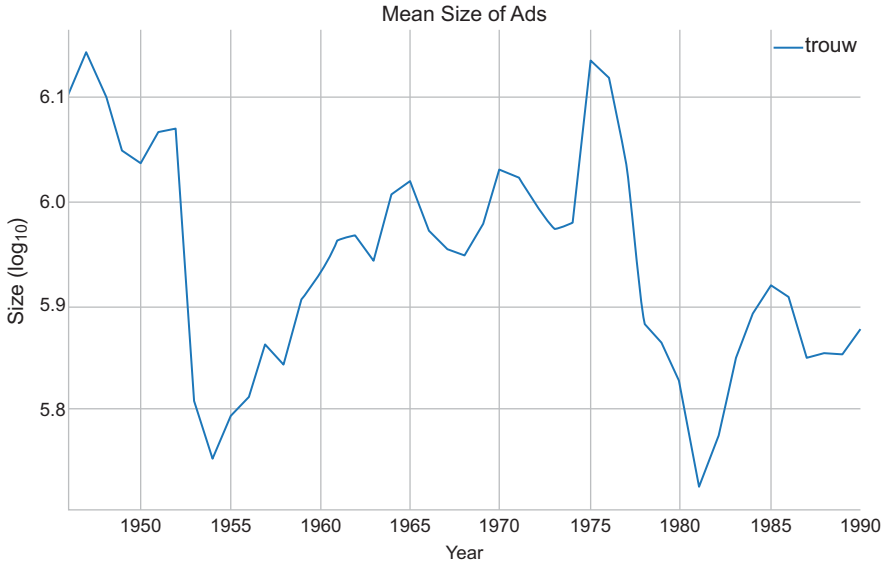
Fig 3: Mean size of advertisements in *Trouw*.

To get a better grasp of the groups of ads based on their size and their position in the newspaper, we can turn to unsupervised clustering methods. These methods can detect groups or clusters within a data set based on a set of features. We base this clustering on the features: width, height, and relative position of the ad in the issue. The latter indicates how close the ad is to the front page or the last page. The distribution of these features contains multiple peaks, indicative of multiple sub-distributions. To be able to capture the distributions of these sub-populations, we use Gaussian Mixture modeling, which can estimate the parameters of these mixtures.[17] Compared to K-means clustering, mixture models incorporate information on the co-variance structure of the data, making it possible to capture clusters with varying distributional shapes. We have applied the clustering separately to even and odd pages, since there seem to be different generative principles at play, as evinced by the heat maps in Fig. 4a and 4b. After estimating the optimal number of clusters for even and odd pages ($n = 8$ and $n = 7$), we fit a Gaussian Mixture model to the data.

---

17 Scikit-Learn offers the Gaussian mixture algorithm for the estimation of a mixture model.
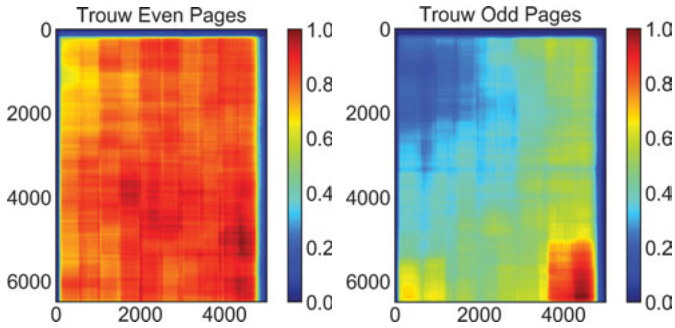
Fig. 4: Heat maps of pixel density in *Trouw*. (a) Pixel Density of Advertisements on Even Pages in *Trouw* (b) Pixel Density of Advertisements on Odd Pages in *Trouw*.

The results of this clustering are presented in Fig. 5a and 5b. These plots are based on a sample of 50,000 ads from even and odd pages. We see apparent differences in size and position in the newspaper between the odd and even pages. The odd pages are less structured, lack a clear signal, and show more noise, especially for ads with larger sizes. The ads on the even pages seem, by contrast, to be more structured. For some of these clusters, there is more variation, while others are closely clustered. A noteworthy cluster on the even pages is the brown cluster, which refers to full-page ads. We can also see that these ads appear toward the end of the newspaper. Newspapers are generally structured in columns, which is visible in the distinct structure present in the widths of advertisements (x-axis). The height of ads (y-axis) has more variance than the width, nevertheless, the algorithm distinguishes clusters based on their height. In terms of relative position, the number of advertisements increases towards the end of the newspaper issue (z-axis). Ads with a small width and a more considerable height, possibly classifieds ads (see Fig. 1 for an example), also appear toward the end of the paper. The information contained in these clusters can help us focus on or filter out particular subsets of advertisements. This method can help us determine the boundaries of these ads in terms of size and position.

We can also engineer other means to help us refine our subsets even further. A useful feature, for instance, is character proportion, which can be calculated by dividing the number of characters by the advert's size (width × height) of the adverts. We can use this metric to filter out textual advertisements or those that predominantly consist of visual material. A similar feature is the ratio of digits to characters. This feature can be used to distinguish ads that mostly list prices of imported goods or wholesale goods, from other ads. These numbers are dependent on the quality of the text extracted by OCR. Notwithstanding variations between newspapers and differences over time, these metrics allow us to filter
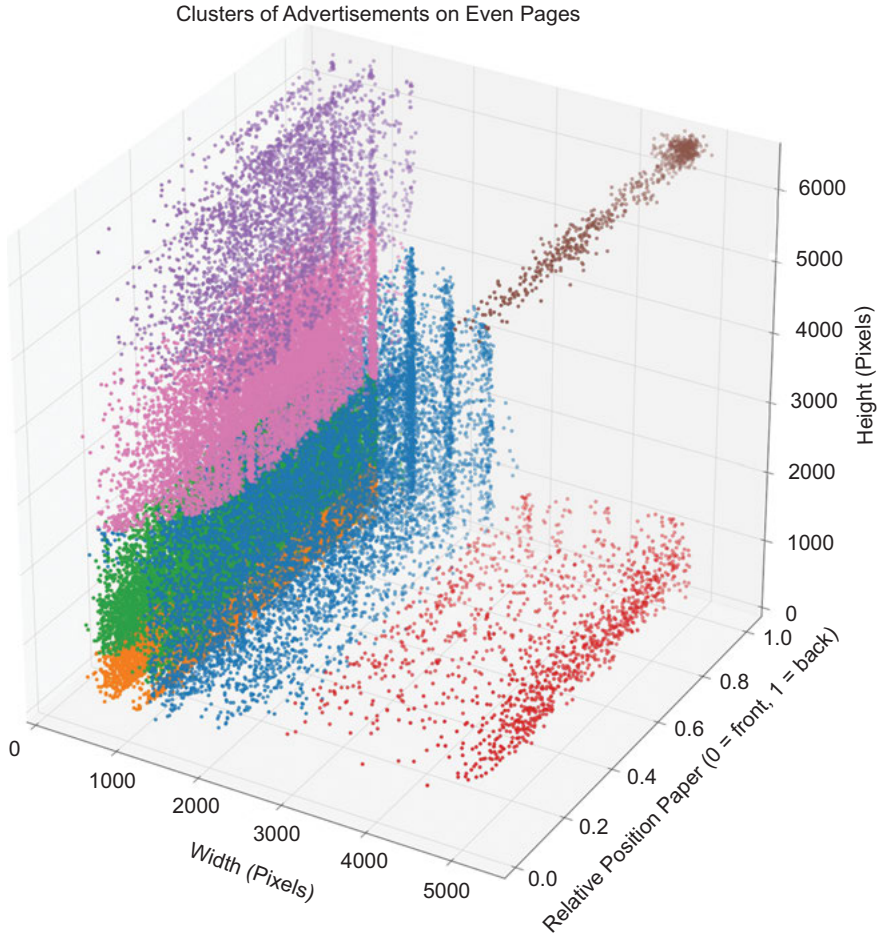
Fig. 5: Clustering of advertisements based on width, height, and relative position in *Trouw*.
(a) Even pages (b) Odd pages.

particular types of advertisements. For example, we can filter out ads by inspecting particular clusters in the distribution of the character proportion. This form of exploratory data analysis can, for example, help to quickly filter out specific types of advertisements, such as classifieds. These advertisements are characterized by a higher than average height and smaller width, as well as a higher character proportion. We could also combine this metadata information with keyword searches to further filter out particular types of ads for selected products.

Clusters of Advertisements on Odd Pages
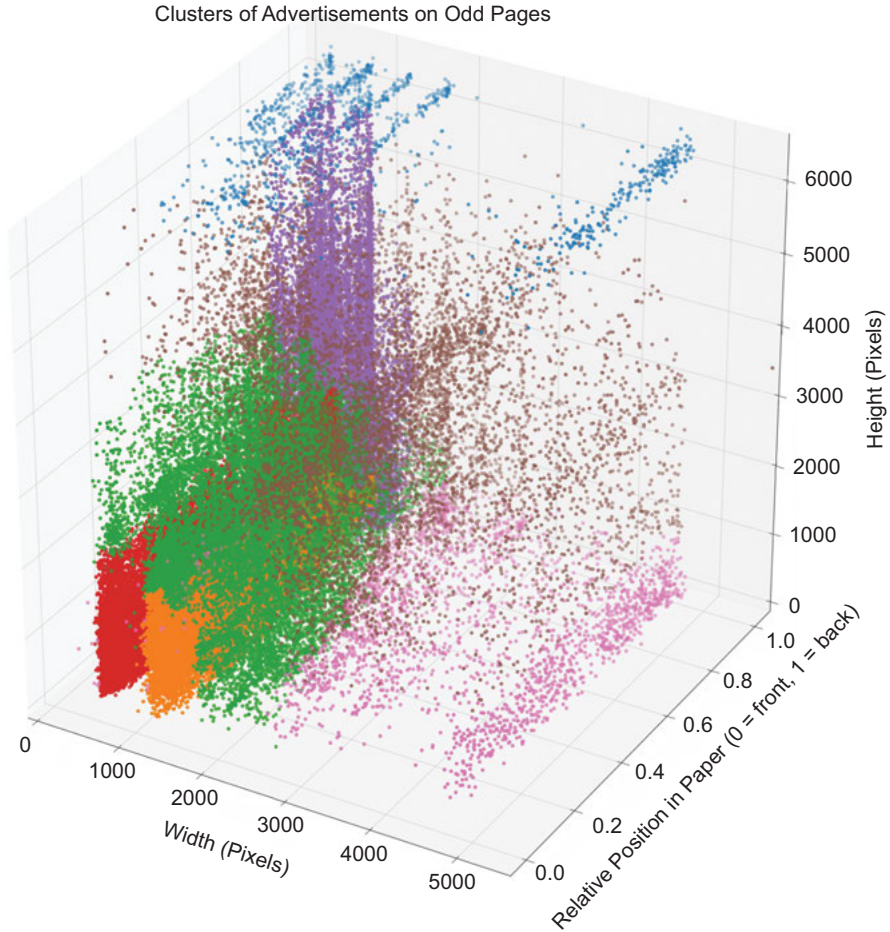

Fig. 5 (continued)

In this section, we have demonstrated a possible approach to using metadata to learn more about the structure and the organization of advertisements in historical newspapers. There are many more directions to explore. However, this section underscores that metadata should not be overlooked, and can be used in conjunction with text analysis to support further analysis.

3 Text Mining Advertisements

In this section, we show how we can extract textual information from historical advertisements, and how this textual information can be used for the analysis of trends and particularities. More specifically, we present a case study on the advertised nationalities of cigarettes. We constructed our corpus of cigarette advertisements from a larger corpus of advertisements from ten national newspapers spanning from 1890 to 1990 (Tab. 1).[18]

Tab. 1: Overview of selected national newspapers.

Newspaper	Period	Number of ads
Algemeen Handelsblad	1906–1970	979,312
Het Volk	1919–1945	191,626
Nieuwe Rotterdamsche Courant	1909–1929	472,536
NRC Handelsblad	1970–1990	460,996
Parool	1945–1990	1,626,204
Telegraaf	1893–1990	3,777,982
Trouw	1946–1990	1,154,746
Vaderland	1919–1945	317,440
Volkskrant	1940–1995	1,193,558
Vrije Volk	1945–1990	1,584,863

From the large corpus of advertisements, we extracted a sub-corpus of advertisements that contain the three most common spelling variants of cigarette: 'cigaret', 'sigaret', and 'cigarette'. These singular and plural variants are queried using the following regular expression: 'cigaret*\w+', 'sigaret*\w+'. This query yielded 43,781 advertisements. Figure 6 displays the distribution of the relative number of these cigarette advertisements. The trend line shows that the relative number of cigarette advertisements grew until the early 1920s, after which it decreased. In the 1950s, the relative number of ads again peaked, and after the 1960s, the number of advertisements that included variants of 'cigarette' dropped considerably, suggesting that after the 1960s, cigarette manufacturers advertised less in Dutch newspapers.

Plotting time series gives an overview of the temporal distribution of advertisements for a particular product. However, the search terms should be carefully selected. Using the different spelling variations, captured with the

---

18 Not all of these ten newspapers appeared throughout the entire period. Advertisements were selected using the metadata field article type.
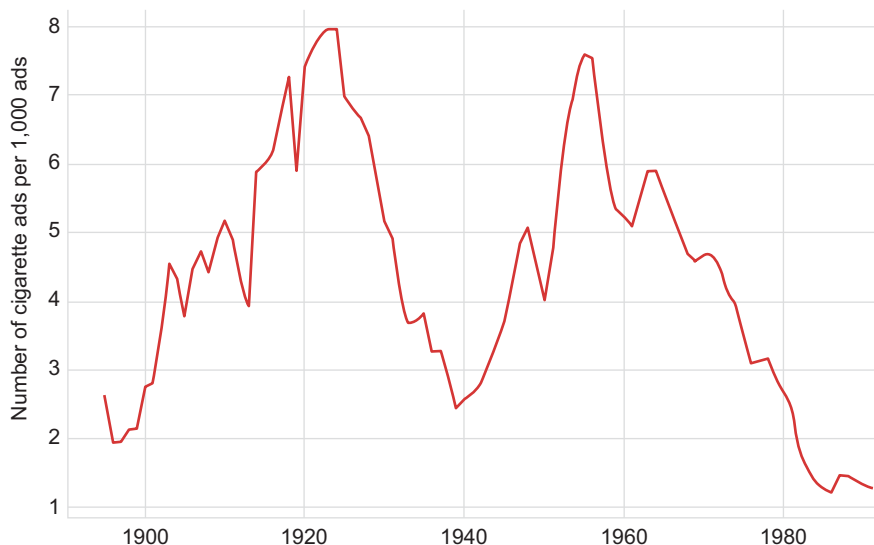
Fig. 6: Relative Number of Cigarette Advertisements 1890–1990.

regular expression, this query covers most of the historical variations in the twentieth century. In the case of cigarettes, the decrease in the last quarter of the twentieth century can in part be explained by the fact that advertisers stopped referencing cigarettes explicitly. Rather, they turned to ads that include brand names or logos, but not 'cigarette'. In such instances, computer vision could help to detect depictions of cigarettes or the occurrence of particular logos in advertisements. Many of these brand names are ambiguous, and querying them could also return advertisements for unrelated products. Moreover, the period during the Second World War should also be scrutinized more closely, as the newspaper landscape differs considerably, while the size and content of newspapers transformed drastically. Even when using relative frequencies to offset the changing number of newspapers, we cannot account for such factors.

3.1 Cigarettes and Product Nationalities

While text mining enables the study of large-scale trends, it can also help to explore more specific ways in which advertisers framed a product. One type of framing is the nationality of a consumer good, also known as 'product nationality'. The relationship between products and constructions of nationality has long

since been debated in consumer and marketing studies.[19] A key finding in these studies is that a strong connection to a favored country persuades consumers to spend more money on products with the same geographical connection.[20] The representations of product nationality are more than mere associations, they are "powerful narratives about the meanings and values transferred by products from their origin to their destination".[21]

In the case of the cigarette, the representation of product nationalities is an under-researched topic, which is surprising given that advertisers explicitly connected cigarettes to particular countries and regions. These product nationalities usually referred to their actual or perceived country of origin, but often also communicated particular characteristics associated with these countries.

The advertisement in Fig. 7 is a clear example of the association between product nationality and particular characteristics. This 1938 advert for the 'Buffalo' cigarette linked the brand to the United States in several ways. First, the brand name 'Buffalo' denoted the emblematic American prairie animal, as well as the city of Buffalo in upstate New York. This relationship between the brand and the United States was further enforced by a small print mentioning its producer: The Cumberland Company from Clarksville, USA. Secondly, in addition to these textual cues, the advertisement included a visual signifier: a background image of a giant cowboy bending over a Dutch tulip field. This picture of a cowboy – an exemplar of American culture – further substantiated Buffalo as an American cigarette. Furthermore, this image expressed the towering dominance of American products in the Netherlands. Thirdly, the ad presented the Buffalo cigarette as having an American product nationality by describing it as "the tastiest and spiciest American cigarette." The geographical association to the United States suggested the product's country of origin, but also signified a particular taste specific to American techniques of tobacco preparation.

---

19 Mrugank Thakor (1996). "Brand Origin: Conceptualization and Review." In: *Journal of Consumer Marketing* 13.3, pp. 27–42; Gordon Hull (Mar. 2016). "Cultural Branding, Geographic Source Indicators and Commodification." In: *Theory, Culture & Society* 33.2, pp. 125–145.

20 Luisa Menapace et al. (2011). "Consumers' Preferences for Geographical Origin Labels: Evidence from the Canadian Olive Oil Market." In: *European Review of Agricultural Economics* 38.2, pp. 193–212.

21 Søren Åskegaard and Güliz Ger (1998). "Product-Country Images: Towards a Contextualized Approach." In: *European advances in consumer research* 3.1, pp. 50–58.

Fig. 7: Advertisements for Buffalo cigarettes in *Limburger Koerier*, April 13, 1938.

The Buffalo ad is, of course, just one example. How does this particular advertisement compare to others, and is its messaging specific to cigarettes with an American product nationality? Using text mining, we can gather the information that helps us answer such questions.

3.2 Charting Product Nationalities

We can, for example, chart the nationalities most commonly associated with cigarettes in advertisements. There are two basic ways to establish product nationality. First, we could count the occurrences of bi-grams – two adjacent word tokens – that include an explicit reference to nationality and cigarettes. For example, to find out when and how often advertisers described cigarettes as American, one could count the advertisements that contained bi-grams such as 'American Cigarette' (*Amerikaanse sigaret*). Regular expressions can help to capture a wide range of possible spelling variations of such a bi-gram. However, a key flaw in counting the bi-gram 'American cigarette' is that we then only enumerate instances in which 'American' appeared directly to the left of 'cigarette'. Advertisers, however, also used other ways to relate cigarettes to a particular location. For instance, in the case of Egyptian cigarettes, advertisers relied on phrases such as "imported from Egypt." One possible yet time-consuming solution is to construct a list of possible strings that express a relationship to a specific nationality.

A second method is to count references to nationality, *i.e.* country names, that co-occur in the proximity of 'cigarette'. In this approach, one only counts words that co-occur within a specific span of words, rather than advertisements that contain the two words. The co-occurrence of words in one advertisement does not necessarily indicate a relationship between them. 'American' could appear in an advertisement for cigarettes without referring to the product nationality of the cigarette or one of its features. Nevertheless, word proximity is a good indicator of a semantic relationship between the given words. Hence, in what follows, we count references to America within a span of five words to the right or left of the keyword 'cigarette'.

An added benefit of proximity searches is that it helps to mitigate issues related to incorrect document segmentation. During digitization, the Optical Layout Recognition (OLR) did not always correctly segment advertisements. There are, for instance, cases where 'cigarette' and 'America' appeared in the actual newspapers in two separate advertisements, whereas after digitization, they were identified as one single advertisement (see Fig. 8). In Fig. 8, 'American' refers to the American hotel in Amsterdam, and 'cigarettes' *sigaretten* to an unassociated retailer advertising the product. The words appeared in one single advertisement, albeit separated by a large number of words. Therefore, the use of a span of five words would not have counted this as an instance in which the two words appeared together. Looking for words in proximity to each other reduces the impact of errors produced by composite advertisements.

Fig. 8: Example of an incorrect segmentation of advertisements. *Algemeen Handelsblad*, January 15, 1883.

For this section, we charted six product nationalities associated with cigarettes: American, British, Egyptian, Russian, Turkish, and Virginia.[22] Using regular expressions, we queried the singular and plural variants of these references as well as common spelling variations. The occurrence of these words in advertisements serves as a proxy for the popularity of cigarettes with varying product nationalities. Figure 9 shows the relative frequency of references to nationalities per 1,000 cigarette advertisements.

From 1890 to 1919, Egyptian, Turkish, and Russian cigarettes were the most popular. The popularity of these cigarettes mirrored the economic, cultural, and political power of the associated geopolitical entities. Before the First World War, the popularity of Russian and Turkish cigarettes mirrored the might of the Ottoman and Russian Empires. In the same period, the American cigarette industry was making its first forays into the European cigarette market.[23] Just after the First World War, Virginia and Egyptian cigarettes became the most popular cigarettes. After the First World War, the British cigarette industry, especially British American Tobacco (BATCO), played a prominent role in the production and dis-

---

22 We approach Virginia as a nationality since the term came to represent Britain.
23 Allan Brandt (2009). *The Cigarette Century: The Rise, Fall, and Deadly Persistence of the Product That Defined America*. New York: Basic Books.
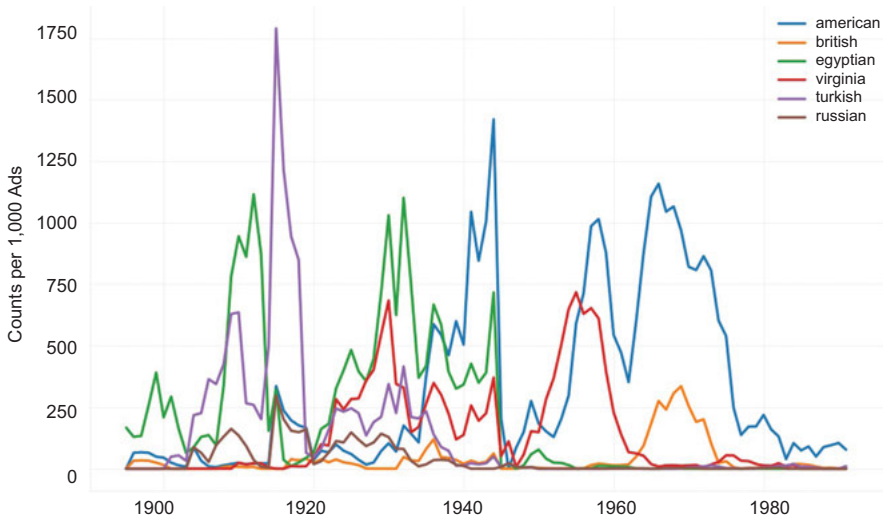
Fig. 9: Relative frequency of signifiers of nationality in cigarette advertisements, 1890–1990.

tribution of both Egyptian and Virginia cigarettes.[24] BATCO was also instrumental in disseminating Virginia cigarettes in the Netherlands.

The association with the United States gained prominence after the Second World War when both American and Virginia cigarettes towered above the other nationalities, which by that time had disappeared almost entirely. In the 1960s, when Virginia cigarettes lost their popularity, the American cigarette acquired sole dominance. A decrease in references to the United States characterized the subsequent decades.[25] This decline coincided with the growing sentiments of anti-Americanism in the Netherlands. Amid anti-American sentiments, advertisers might have refrained from associating their product with the United States. Furthermore, in 1964, the American Surgeon General Luther Terry published the Report on Smoking and Health, in which he presented the detrimental effects

24 Relli Shechter (2006). *Smoking, Culture and Economy in the Middle East: The Egyptian Tobacco Market 1850–2000*. London: Tauris, pp. 27–8.

25 Rob Kroes (2006). "European Anti-Americanism: What's New?" In: *Journal of American History* 93.2, pp. 417–432; Jessica C. E. Gienow-Hecht (Jan. 2006). "Always Blame the Americans: Anti-Americanism in Europe in the Twentieth Century." In: *The American Historical Review* 111.4, pp. 1067–1091.

smoking could have on one's health. The report led to a significant decrease in cigarette consumption in the Netherlands and the United States.[26]

As this section has shown, counting specific strings of text in digitized material is a relatively easy and fast way to gauge and compare the popularity of particular products. The trends in popularity in Dutch newspapers matched global trends. However, the interest in American cigarettes already declined before the Second World War, while academic literature on Americanization in the Netherlands commonly situates this process after the Second World War.

We will now examine how cigarettes with different product nationalities were presented in advertisements. In other words, were there distinct product features for cigarettes with particular product nationalities? Understanding the historical use of product features in advertising discourse can help us determine the cultural and technological impact of such products.

3.3 Product Features

In this section, we discuss two methods to examine the historical evolution of particular characteristics associated with cigarettes. First, we show how an unsupervised method that can detect noteworthy patterns, or 'bursts', in word use can be leveraged to extract words that typified a certain period. Second, we demonstrate how machine learning can determine whether cigarettes with a clear product nationality possessed features distinct from cigarettes with other nationalities.

3.3.1 Finding Trending Topics

When we want to study the use of particular words and examine how this use has evolved, one of the first tasks is to decide which words we are studying. One method is to rely on secondary sources to compile a vocabulary of words related to a subject or a particular period. A different approach is a data-driven one, in which we extract words that exhibit noteworthy use in the corpus using algorithms. One such algorithm is 'burst detection'.

---

26 Dietrich Hoffmann, Ilse Hoffmann, and Karam El-Bayoumy (2001). "The Less Harmful Cigarette: A Controversial Issue. A Tribute to Ernst L. Wynder." In: *Chemical research in toxicology* 14.7, pp. 767–790.

Burst detection is a modeling technique to detect *bursts of activity* in streams of data, for example, the sudden rise and fall in word frequency in serial publications. The trends for individual words cannot easily be compared. For example, for words with relatively little overall activity, a sudden, repetitive increase in use can signal a burst, whereas, for words with much activity, a different burst intensity might be required. Moreover, bursts of activity can also be nested within larger patterns. To be able to capture these bursts, Jon Kleinberg developed an algorithm that models the stream of information as an infinite-state automaton. The algorithm assigns costs to state transitions, which makes it possible to distinguish between short bursts and long burst even while the overall rate of transmission changes over time.[27] We use Kleinberg's algorithm to detect whether and when specific words exhibited "bursty" behavior in advertising discourse. In other words, we use it to identify *trending topics* in advertising discourse.

We model the 'burstiness' for a subset of words. This subset includes the 500 most distinctive determined adjectives and nouns – determined using tf-idf – from the corpus of cigarette advertisements. We then apply Kleinberg's algorithm using the default settings on monthly frequency counts of these 500 words. For each of these words, we get information on whether they bursted and how intense the burst was, as well as the duration of the burst. Figure 10 displays the top 50 bursty nouns and adjectives and when they bursted. We can, for example, see the appearance of the key term 'filter' around 1965, when the Surgeon General's report also appeared.

Moreover, the debate in the late 60s and early 1970s shifted toward the amount of nicotine in cigarettes. The appearance of the term 'health' (*gezondheid*) in the mid-1920s stems from advertisements that promoted the health benefits of smoking Virginia cigarettes. The figure also contains phrases such as 'job application' (*sollicitatie*) and 'human resources' (*personeelszaken*). Upon closer inspection, these words appeared in job ads placed by cigarette companies, a category that researchers might want to prune from their corpus.

Using this method, we can quickly gauge when particular words displayed 'bursty' behavior or which topics were trending at particular moments in time. This information can subsequently be used for closer examination of the texts in which these words appeared. Such techniques can also help distinguish a particular subset that one might want to remove, or treat as a separate corpus.

---

27 Jon Kleinberg (2002). "Bursty and Hierarchical Structure in Streams ⋆." In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 25.

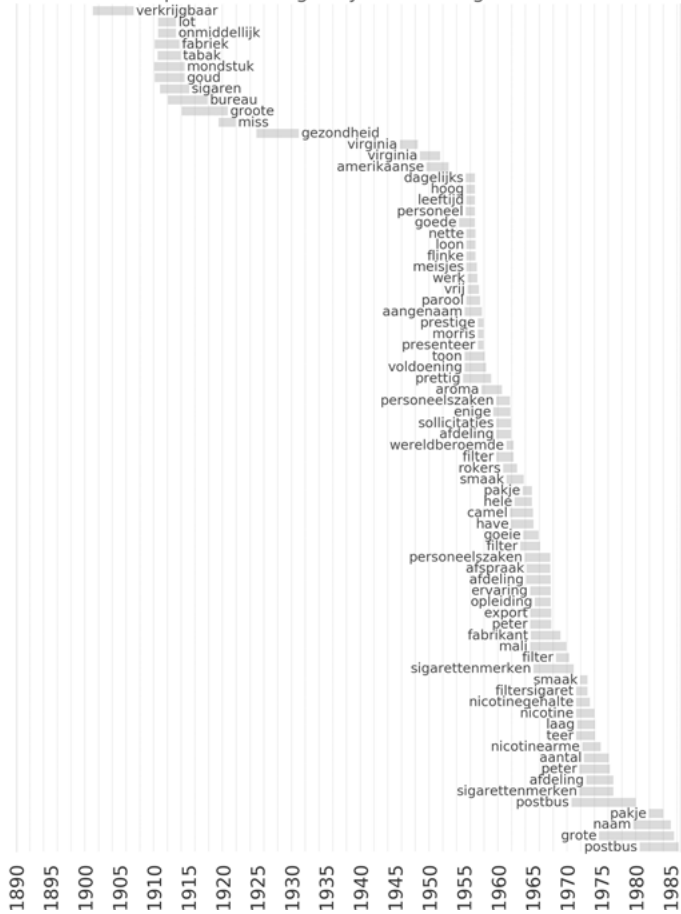Timeline of the top 75 "bursting" keywords in Cigarette Advertisements

verkrijgbaar
lot
onmiddellijk
fabriek
tabak
mondstuk
goud
sigaren
bureau
groote
miss
gezondheid
virginia
virginia
amerikaanse
dagelijks
hoog
leeftijd
personeel
goede
nette
loon
flinke
meisjes
werk
vrij
parool
aangenaam
prestige
morris
presenteer
toon
voldoening
prettig
aroma
personeelszaken
enige
sollicitaties
afdeling
wereldberoemde
filter
rokers
smaak
pakje
hele
camel
have
goeie
filter
personeelszaken
afspraak
afdeling
ervaring
opleiding
export
peter
fabrikant
mali
filter
sigarettenmerken
smaak
filtersigaret
nicotinegehalte
nicotine
laag
teer
nicotinearme
aantal
peter
afdeling
sigarettenmerken
postbus
pakje
naam
grote
postbus

1890 1895 1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960 1965 1970 1975 1980 1985

Fig. 10: Top 50 Bursty Nouns and Adjectives in Cigarette advertisements, 1890–1990.

3.3.2 Detecting Distinctive Features

A different line of questioning might focus on the differences in vocabulary between cigarettes with varying product nationalities. For example, were American cigarettes more commonly associated with a mild taste, or were Egyptian cigarettes presented as artisanal? One method to gauge the differences between texts is through the use of a machine learning classifier. These classifiers learn which textual features are predictive of a certain class of texts. For example, which words are indicative of either American cigarettes or Virginia cigarettes.

This method assumes that language use between the two remains separated over time. In other words, if features are first introduced in Virginia cigarettes and then co-opted by American cigarettes, it might be more challenging to use this feature to distinguish between the two. For this reason, we can also train classifiers for separate periods, given that we have enough data. If the classifiers fails to separate the two corpora, we can infer that the differences between language use were not very large. For each of these periods, we can also investigate the most informative features, *i.e.* the features that the algorithms rely the most on to make the classification.

We only select nouns, adjectives, adverbs, and verbs in the text.[28] Next, we label the corpus as either American or British, based on whether the ads contain references to the United States or to Virginia and the United Kingdom. After removing explicit references to these nationalities, we train a Naive Bayes Classifier on the data.[29] Based on the occurrence of particular words, the classifier learns to predict whether a text is an advertisement for American or British cigarettes. For the period 1920–1980, in which both types of cigarettes were represented, the classifier can distinguish between advertisements for American and Virginia cigarettes with an accuracy of 0.88, and respective F_1-scores of .92 and 0.79.

Subsequently, we can examine the most informative features for the classifier to label an ad as either American or Virginia. Noteworthy results include brand names such as 'Lucky Strike', 'Roxy', and 'Camel' for American cigarettes, and 'Derby' and 'Chief Whip' for Virginia cigarettes. Moreover, words such as 'cork' (*kurk*), 'mouth piece' (*mondstuk*), 'purity' (*zuiverheid*), and 'health' (*gezondheid*) were predictive for Virginia cigarettes, while 'connaisseurs' (*liefhebbers*), 'packages' (*pakjes*), 'filter', and 'smoking pleasure' (*rookgenot*) are all predictive of American cigarettes.

There was one company primarily responsible for connecting these features to Virginia cigarettes: 'Ardath'. This British cigarette manufacturer boasted that the purity of its Virginia tobacco led to a better tasting and healthier cigarette. This link was particularly strong in advertisements for the brand Chief Whip, which Ardath described as the "zenith of purity" (*toppunt van puurheid*). Together with the company Wills, Ardath distinguished the taste of Virginia cigarettes from that of American cigarettes. Ardath and Wills both denounced saucing and blending – two key features of the American cigarettes – and distanced themselves from American cigarettes in doing so. In 1925, Wills claimed

---

28 For this, we rely on Spacy. https://spacy.io.

29 Without removing these explicit references, it would be very easy for the classifier to determine in which category a text would fall. We use the NLTK implementation of the Naive Bayes Classifier (https://www.nltk.org/_modules/nltk/classify/naivebayes.html).

that its Virginia cigarettes consisted of 100% pure Virginia tobacco, without the addition of Greek, Indonesian, or Turkish tobaccos. In advertisements for Chief Whip, a doctor claimed that the cigarette was "absolutely pure and free of all surrogates and sauces." Advertisers presented Chief Whip as a pure and unprocessed cigarette.

The features used by the classifier point to three distinctive material aspects of the American cigarette: its length, its filter tip, and its packaging. Three additions that American cigarette producers introduced in Europe after the Second World War. First, in the 1950s, the longer, king size cigarette was introduced to Dutch consumers, after enjoying great success in the United States. Advertisers linked the increase in length to the United States to help familiarize the Dutch smoker with the long cigarette. In a 1955 advertisement, the brand So Long referred to the United States in their explanation of a king size cigarette: "In America, a cigarette longer than 85mm is called King Size."[30] Advertisers used the link with the United States to help acquaint Dutch consumers with longer cigarettes.

The second significant change was the introduction of filter cigarettes. Amid growing health concerns in the United States, American cigarette manufacturers introduced the purportedly healthier filter cigarettes in the early fifties. In the context of American filter cigarettes, purity did not denote the unblended nature of cigarettes, but the purifying effects of filters. Advertisements claimed that a lengthy filter would lead to "more and purer smoking."[31] Filter cigarettes became hugely popular, and by the 1970s, almost ninety percent of the cigarette market consisted of filter cigarettes.[32]

Finally, technological developments in the United States changed the look and feel of cigarette packaging. These innovations offered advertisers new ways to link the product to the United States. American companies were the first to package cigarettes mechanically in plastic-wrapped cardboard boxes. Until then, consumers bought cigarettes per piece, and stored them in less practical tin boxes. The new method of packaging not only referred to the United States because of its origin, but it also carried cultural connotations that resonated with American culture. Advertisers described the packaging of the American cigarette as flat, practical, modern, famous, or fancy.

---

30 "So Long advertisement," *Het Vrije Volk*, December 13, 1955.

31 "Sir Richard advertisement," *De Telegraaf*, January 5, 1962.

32 Allan Brandt (2009). *The Cigarette Century: The Rise, Fall, and Deadly Persistence of the Product That Defined America*. New York: Basic Books, p. 244.

The information gained through this method can be used for closer examinations of particular characteristics, as demonstrated above. A more in-depth analysis falls outside of the scope of this chapter.[33]

4 Conclusion

Advertisements are a rich and varied source for the study of public discourse. This chapter showcases how we can apply computational methods to metadata information to learn more about advertising in the past. The ability to distinguish between different types of advertisements can help to streamline further inquiry. Moreover, based on a case study on product nationalities in cigarette advertisements, we showed how text mining can help to better understand the historical trajectories between products and particular nationalities. Additionally, we demonstrated how machine learning can help to identify trending topics and distinctive words in advertising discourse.

There is always an interplay between prior knowledge, the choice and implementation of the computational methods, and the interpretation of results. Therefore, it is important that we are explicit about the assumptions that we introduce into our computational modelling. This prevents us from massaging the data or optimizing the method to merely find what we want to find. As Joshua Epstein succinctly shows, models can help us with more than prediction and can help us with the data collection, explanation of data, and the surfacing of dynamics in data.[34] While the methods introduced in this article can certainly be applied to different domains, they always need to be fine-tuned to the data and specific question. Moreover, relying on a combination of methods, including ones that show uncertainty in results, can further improve the transparency and the value of computational methods in historical research. While the text quality is often far from perfect in digitized newspapers, there is still much we can extract from this abundance of source material using computational methods. With the increasing amount of digitized newspaper collections in different countries, we can also envision expanding our efforts to study advertisements from a transnational perspective.

---

33 See Melvin Wevers (2017). "Consuming America: The United States as a Reference Culture in Dutch Public Discourse on Consumer Goods, 1890–1990." Ph.D. Dissertation. Utrecht: Utrecht University, for such an analysis.

34 Joshua M. Epstein (Oct. 2008). "Why Model?" In: *Journal of Artificial Societies and Social Simulation* 11.4, p. 12.

Bibliography

Arnold, Taylor and Lauren Tilton (Mar. 2019). "Distant viewing: analyzing large visual corpora." In: *Digital Scholarship in the Humanities*. DOI: 10.1093/digitalsh/fqz013. eprint: https://academic.oup.com/dsh/advance-article-pdf/doi/10.1093/digitalsh/fqz013/28082598/fqz013.pdf. URL:https://doi.org/10.1093/digitalsh/fqz013.

Åskegaard, Søren and Güliz Ger (1998). "Product-Country Images: Towards a Contextualized Approach." In: *European advances in consumer research* 3.1, pp. 50–58.

Barman, Raphaël, Maud Ehrmann, Simon Clematide, Sofia Ares Oliveira, and Frédéric Kaplan (Jan. 2021). "Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers." In: *Journal of Data Mining & Digital Humanities* HistoInformatics. URL:https://jdmdh.episciences.org/7097.

Bilgin, Aysenur, Laura Hollink, Jacco van Ossenbruggen, Erik Tjong Kim Sang, Kim Smeenk, Frank Harbers, and Marcel Broersma (Oct. 2018). "Utilizing a transparency-driven environment toward trusted automatic genre classification: A case study in journalism history." In: *IEEE 14th International Conference on eScience, e-Science 2018*, pp. 486–496. DOI: 10.1109/eScience.2018.00137.

Brandt, Allan (2009). *The Cigarette Century: The Rise, Fall, and Deadly Persistence of the Product That Defined America*. New York: Basic Books.

Chung, Joon Son, Relja Arandjelović, Giles Bergel, Alexandra Franklin, and Andrew Zisser- man (2015). "Re-Presentations of Art Collections." In: *Computer Vision – ECCV 2014 Workshops*. Ed. by Lourdes Agapito, Michael M. Bronstein, and Carsten Rother. Cham: Springer International Publishing, pp. 85–100.

Daems, Joke, Thomas D'haeninck, Simon Hengchen, Tecle Zere, and Christophe Verbruggen (June 2019). "'Workers of the World'? A Digital Approach to Classify the Interna- tional Scope of Belgian Socialist Newspapers, 1885–1940." In: *Journal of European Periodical Studies* 4.1, pp. 99–114. DOI: 10.21825/jeps.v4i1.10187.

Epstein, Joshua M. (Oct. 2008). "Why Model?" In: *Journal of Artificial Societies and Social Simulation* 11.4, p. 12.

Fox, Stephen R. (1997). *The Mirror Makers: A History of American Advertising and Its Creators*. English. Urbana: University of Illinois Press.

Gienow-Hecht, Jessica C. E. (Jan. 2006). "Always Blame the Americans: Anti-Americanism in Europe in the Twentieth Century." In: *The American Historical Review* 111.4, pp. 1067–1091.

Hill, Mark J. and Simon Hengchen (Dec. 2019). "Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study." In: *Digital Scholarship in the Humanities* 34.4, pp. 825–843.

Hoffmann, Dietrich, Ilse Hoffmann, and Karam El-Bayoumy (2001). "The Less Harmful Cigarette: A Controversial Issue. A Tribute to Ernst L. Wynder." In: *Chemical research in toxicology* 14.7, pp. 767–790.

Hull, Gordon (Mar. 2016). "Cultural Branding, Geographic Source Indicators and Com- modification." In: *Theory, Culture & Society* 33.2, pp. 125–145.

Kleinberg, Jon (2002). "Bursty and Hierarchical Structure in Streams." In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 25.

Kroes, Rob (2006). "European Anti-Americanism: What's New?" In: *Journal of American History* 93.2, pp. 417–432.

Lears, T. Jackson (1994). *Fables of Abundance: A Cultural History of Advertising in America*. New York: Basic Books.

Lee, Benjamin Charles Germain, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S. Weld (2020). "The Newspaper Navigator Dataset: Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling America." In: *Proceedings of the 29th ACM International Conference on Information Knowledge Management*. CIKM '20. Virtual Event, Ireland: Association for Computing Machinery, pp. 3055–3062. DOI: 10.1145/3340531.3412767. URL:https://doi.org/10.1145/33405313412767.

Liebl, Bernhard and Manuel Burghardt (2020). "From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline." In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, The Netherlands, November 18-20, 2020*. Ed. by Folgert Karsdorp, Barbara McGillivray, Adina Nerghes, and Melvin Wevers. Vol. 2723. CEUR Workshop Proceedings. CEUR-WS.org, pp. 351–373. url:http://ceur-ws.org/Vol-2723/long20.pdf.

Marchand, Roland (1985). *Advertising the American Dream: Making Way for Modernity, 1920–1940*. Berkeley: University of California Press.

Marshall, Margaret (1995). *Contesting Cultural Rhetorics: Public Discourse and Education, 1890–1900*. Ann Arbor: University of Michigan Press.

Menapace, Luisa, Gregory Colson, Carola Grebitus, and Maria Facendola (2011). "Consumers' Preferences for Geographical Origin Labels: Evidence from the Canadian Olive Oil Market." In: *European Review of Agricultural Economics* 38.2, pp. 193–212.

Moretti, Franco (2013). *Distant Reading*. London: Verso.

Postman, Neil (Dec. 2005). *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. London: Penguin.

Schudson, Michael (1982). *The Power of News*. Cambridge: Harvard University Press.

Shechter, Relli (2006). *Smoking, Culture and Economy in the Middle East: The Egyptian Tobacco Market 1850–2000*. London: Tauris.

Thakor, Mrugank (1996). "Brand Origin: Conceptualization and Review." In: *Journal of Consumer Marketing* 13.3, pp. 27–42.

van Eijnatten, J. and Ruben Ros (2019). "The Eurocentric Fallacy: A Digital-Historical Approach to the Concepts of 'Modernity', 'Civilization' and 'Europe' (1840–1990)." In: *International Journal for History, Culture and Modernity* 7. doi: https://doi.org/10.18352/hcm.580

van Strien, Daniel, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza (2020). "Assessing the Impact of OCR Quality on Downstream NLP Tasks:" In: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*. Malta: SCITEPRESS, pp. 484–496.

Vree, Frank van (1989). *De Nederlandse pers en Duitsland 1930–1939: een studie over de vorming van de publieke opinie*. Groningen: Historische Uitgeverij.

Wevers, Melvin (2017). "Consuming America: The United States as a Reference Culture in Dutch Public Discourse on Consumer Goods, 1890–1990." Ph.D. Dissertation. Utrecht: Utrecht University.

Wevers, Melvin and Thomas Smits (Apr. 2020). "The Visual Digital Turn: Using Neural Networks to Study Historical Images." In: *Digital Scholarship in the Humanities* 35.1, pp. 194–207. doi: 10.1093/llc/fqy085.